(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
23 October 2003 (23.10.2003)

PCT

(10) International Publication Number
WO 03/087837 A2

(51) International Patent Classification[7]: G01N 33/68, G06F 19/00

(21) International Application Number: PCT/EP03/03995

(22) International Filing Date: 16 April 2003 (16.04.2003)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/373,162 16 April 2002 (16.04.2002) US
60/410,628 12 September 2002 (12.09.2002) US

(71) Applicant *(for all designated States except US)*: **GENE-PROT, INC.** [CH/CH]; 2, rue Pré de la Fontaine, Case Postale 125, CH-1217 Meyrin 2 (CH).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **HOCHSTRASSER, Denis** [CH/CH]; c/o GeneProt, Inc., 2, Pré de la Fontaine, CH-1217 Meyrin (CH). **KUGENER, Veronique** [FR/CH]; c/o GeneProt, Inc., 2, Pré de la Fontaine, CH-1217 Meyrin (CH). **ROSE, Keith** [GB/CH]; c/o GeneProt, Inc., 2, Pré de la Fontaine, CH-1217 Meyrin (CH).

(74) Agent: **VOSSIUS & PARTNER**; Siebertstrasse 4, 81675 Munich (DE).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 03/087837 A2

(54) Title: METHODS AND SYSTEMS FOR DONOR SAMPLE MANAGEMENT AND POOLING FOR USE IN PROTEOMIC STUDIES

(57) Abstract: The present invention relates to methods and systems for identifying subjects from whom biological samples may be obtained and used in proteomic studies. More specifically, the present application relates to methods for obtaining biological samples having preselected criteria, such as a chosen level of protein quality. The present application relates as well to methods for preparing pools of biological samples.

METHODS AND SYSTEMS FOR DONOR SAMPLE MANAGEMENT AND POOLING FOR USE IN PROTEOMIC STUDIES

FIELD OF THE INVENTION

The present invention relates to methods and systems for identifying subjects from whom biological samples may be obtained and used in studies including, but not limited to, biomarker discovery, target discovery, protein therapeutics discovery and monitoring the effectiveness of disease treatment.

More specifically, the present application relates to methods for obtaining biological samples having preselected criteria, such as, by way of example and not limitation, a chosen level of protein quality. The present application relates as well to methods for preparing pools of biological samples. In preferred embodiments, the invention provides means for preparing pooled plasma samples.

BACKGROUND OF THE INVENTION

The discovery of candidate proteins as biomarkers for diagnostics, targets for therapeutic intervention or as protein therapeutics has been accelerated through recent developments in proteomics technologies. These technologies, however, are limited in their ability to detect the full complement of proteins in a sample (also known as "proteome"), particularly low-abundance proteins, especially given the complexity of the proteome from a mammalian biological sample such as, e.g., plasma. For example, it is well known that the activity of many proteins is modulated by posttranslational modifications, such as phosphorylation, glycosylation or proteolytic cleavage and these modifications increase the proteome complexity.

The analysis of a proteome involves the separation of the proteins in a sample followed by the identification of the resolved proteins, a challenging task given the tremendous chemical heterogeneity in virtually all parameters that can be measured.

For example, some cytokines weigh between 1 and 2kD, while some muscle proteins weigh close to 1000kD. Some proteins are very soluble and may also be present at high concentration in aqueous media (e.g. albumin, 40mg/mL in plasma, Putnam, R.W., *The Plasma Proteins*, Academic Press, New York 1975), while some membrane proteins have more than 75% of their amino acids buried in the phospholipid bilayer and are thus not easily amenable to studies in aqueous solvents.

Moreover, proteins are present in extremely divergent concentrations in cells, ranging from 100 to 100,000,000 molecules per cell. This is a major issue for the complete analysis of a proteome, as some high concentration proteins will be masking some of the low abundance ones. This is a particularly difficult problem to resolve, since there are no readily available methods for precisely separating abundant from non-abundant proteins. Current methods generally still result in the unwanted removal of other non-abundant proteins. Some methods have been developed to address this problem, such as the use of specific ligands to remove certain abundant proteins such as albumin or immunoglobulins (see for example International Patent Publications WO 99/65943 and WO99/63351, US Patent 6,410,692 and Lollo BA, Harvey S, Liao J, Stevens AC, Wagenknecht R, Sayen R, Whaley J, Sajjadi FG, Improved two-dimensional gel electrophoresis representation of serum proteins by using ProtoClear. *Electrophoresis*, 1999 Apr-May;20(4-5):854-9, which are incorporated herein in their entirety). For example, by removing one or more abundant proteins from the sample, it is easier to evaluate less abundant proteins as possible disease markers. When the sample is blood derived, e.g. serum or plasma, the abundant proteins typically are immunoglobulin G and albumin. In serum, albumin constitutes about 57-71% of the total serum protein content and immunoglobulin G constitutes 8 -26% of the total serum protein content (Putnam, R.W., *The Plasma Proteins*, Academic Press, New York 1975). Methods described thus far have therefore removed selectively most abundant proteins by using binding proteins, e.g. an antibody or a fragment thereof to remove, e.g., albumin, and Protein A or Protein G to remove Immunoglobulins, all of these preferably immobilized on a solid support.

In view of the complexity of the mammalian proteome, and since the complement of proteins and amounts of each protein in a subject is dependent on a range of environmental and other factors, it is desirable to analyze biological samples from a number of subjects. It may also be desirable to pool the samples prior to the principal fractionation or separation steps, thereby more accurately representing the proteins present in the average subject in the group, and allowing larger volumes to be processed, thereby permitting increased detection sensitivity levels.

The group of subjects for analysis can be selected randomly, but is preferably selected based on their exhibiting a trait such as a medical condition or based on their belonging to a particular population. For example, subjects suffering from a disease can be selected, and the samples there from compared against samples from subjects not suffering from the disease.

It would thus be advantageous to provide a method of selecting subjects for inclusion in a study group, which takes into account the protein quality of the samples to be analyzed and

optionally additional criteria, and which, optionally, provides an optimized way to pool the selected samples for proteomic analysis.

SUMMARY OF THE INVENTION

The present invention provides a method to identify and recruit subjects for participation in proteomics studies and analyses based on medical histories and/or medical data, and/or clinical and/or biological characterization of a donor sample. Biological samples obtained from the donors can then be used in any desired method, for example, to be included in a pooled sample (e.g. as can be used in the comparison of protein abundances between a disease sample and a non-disease sample). In this context, the invention provides methods and systems whose capabilities can be used to identify factors useful for the selection of subjects, e.g. factors that are correlated with aberrant and/or lowered protein sample quality for proteomics use.

The processes contemplated include methods of identifying biological sample donors, methods of identifying biological samples for use in research methods such as proteomics research, methods for selecting or eliminating biological samples or donors in a research method, and methods of performing a proteome analysis. Also envisioned are methods for selecting samples for combination in a pooled sample. Also envisioned is a method of identifying traits that may lead to diminished quality of a biological sample.

In a preferred aspect, the invention relates to a method for selecting biological sample donors, for inclusion in a research method, the method comprising the steps of:

(a) selecting or providing one or more biological sample type(s) to be used in the research method;

(b) optionally selecting or providing one or more fractionation strategy(ies) for the biological sample type(s) selected in (a);

(c) optionally selecting or providing a statistical method for the analysis of the results of the research method and deriving therefrom numbers of biological samples needed;

(d) selecting one or more trait(s), which are preferably clinical or biological criteria, for the assessment of donors, with the *proviso* that at least one of these traits affects the quality of proteins in the donor's biological sample(s);

(e) providing a plurality of sample donors and optionally obtaining one or more biological sample(s) from at least a portion of said plurality of donors;

(f) assessing whether each donor displays the trait(s) selected in step (d);

(g) selecting one or a plurality of donor(s) for inclusion in the study protocol, preferably for inclusion of their respective biological sample(s) in a pooled sample, wherein said donor(s) are selected according to display of the trait(s) assessed in step (f); and

(h) optionally assessing whether the selection from step (g) fulfills the statistical criteria set in step (c) and consequently repeating the method from step (e) onwards, if needed.

Preferably, the method comprises carrying out steps (d), (e), (f) and (g). In other embodiments, the method can be carried out starting from step (d) onwards, or from step (b) onwards, or from step (a) onwards, in any suitable order. Any of the steps, in particular steps (d), (e), (f) and (g) may be repeated.

Preferably, a study protocol for the research method in which the samples are to be included has been designed. The design of a study protocol typically includes a definition of a condition or characteristic under study, and preferably also specifies the number of samples needed for study. In some methods, the study protocol may be designed to catalog the proteins present in a sample. Preferably, a study protocol typically defines one or more conditions or characteristics (e.g., trait or disease) under study and preferably also diagnostic criteria to be used in selecting donors or samples for inclusion in the study. Often, a study involves a trait/non-trait comparison, thereby necessitating one or more control donors or samples. In some instances, no satisfactory selection can be made to obtain sufficient sample size. This may be due to a number of reasons, including lack of subjects meeting the criteria, or reduction in number of donors or samples following removing from consideration of donors or samples of lower sample quality. In these cases it may be necessary to refine one or more criteria used in the assessment of the donors. In some cases the study protocol may provide the possibility to use other, similar conditions or characteristics for selecting donors or samples. To select further donors or samples, the methods of the invention can be repeated. In particular, the method above can preferably be repeated from step (d) onwards, or from step (b) onwards, or from step (a) onwards, or the study protocol for the research method can be modified.

A schematic representation of the steps described above is shown in Figure 1.

Additional traits that can be selected in step (d) above include but are not limited to: age of the donor at the time of sample collection, sex of the donor, medical diagnosis and/or prognosis of the donor for one or more disease(s), medical history of the donor, medical familial history of the donor, and genotype of the donor.

It will be appreciated that samples can be obtained at any desired time in the method. That is, biological samples can be obtained or provided by donors before step (a) in the method above. Alternatively, biological samples from donors to be included in a study protocol can be obtained after step (g) in the method above such that unneeded samples are not obtained unnecessarily.

The assessment of donors as to whether they display a selected trait can be done according to any desired method. Typically, the medical history of a donor is considered or screened for an indication that the donor displays a given trait. It is an object of the present invention to encompass a number of possible combinations for the steps (f) and (g) in the method described above. For example, the assessment of the donors (step (f)) can be conducted by screening for donors who display the selected trait(s), but may also be conducted by screening for donors who do not display the selected trait(s). Where more than one trait is being evaluated, any combination of displaying/not displaying each given trait can be used in the method of the invention. Similarly, the selection (step (g)) can be conducted by retaining a selected group of donors, but may also be conducted by eliminating a selected group of donors. The sample or donor selection method may be computer based, partially computer-based or non-computer based.

In a preferred example of selecting blood, plasma or serum samples in step (a), examples of traits that may be selected against include any conditions involving indication of, or concomitant use of, a drug that could cause impaired liver function or impaired renal function. Further examples include diseases known or expected to cause major alteration in plasma protein composition, including gammapathies, multiple myeloma, Waldenstrom's macrogobulinemia and proteinuria. Further details are provided herein.

The step of assessing whether the donor displays at least one trait affecting the protein quality in the donor's biological sample can be carried out according to any suitable method. In one aspect, a medical or diagnostic criteria is considered. A review of a subject's medical history or a detectable symptom can be considered. In preferred embodiments, the invention comprises conducting an assay for at least one, preferably a plurality of, standard laboratory measures. Preferably assays are conducted for clinical chemistry, complete blood count, coagulation tests and/or serology. Examples of assays useful for determination of impaired liver or renal function are provided below.

In another preferred aspect, the method of the invention uses a database to record information relating to the traits characteristic of each donor, and optionally to record information relating to the biological samples associated thereto.

In preferred embodiments, a proteomics study will require at least 2, 5, 10, 20, 50 or 100 biological samples in order to reduce potential error due to variation in protein levels between the samples. Thus, the invention preferably comprises obtaining a biological sample and/or medical data from at least 2, 5, 10, 20, 50, 100, 200, 500 or 1000 individuals. In other preferred embodiments, the methods comprise combining at least 2, 5, 10, 20, 50, 100, or 200 biological samples in a pooled sample. A pooled sample combines, or is composed of, biological samples from a plurality of individual samples, preferably from a plurality of subjects. Where the method of the invention is used to identify prospective donors, the invention preferably comprises identifying or obtaining samples from at least 2, 5, 10, 20, 50, 100, or 200 individuals.

Generally, step (g) in the method described above comprises identifying at least one donor displaying the selected trait(s). More preferably, a plurality of donors displaying the selected trait(s) are identified.

The selected trait which may affect the quality of a sample, or is known to affect the quality of a sample, is preferably a trait involving aberrant levels of a known protein, compared to an individual who does not display said trait. More preferably, the trait is a trait involving elevated levels of an abundant protein, such as for example elevated levels of immunoglobulins. In another aspect, the selected trait is a trait involving aberrant protein degradation.

In preferred aspect, the biological sample is a fluid sample, most preferably a plasma (or serum) or Cerebro Spinal Fluid (CSF) sample. In further aspect the methods of the invention comprise obtaining a biological sample, preferably a plasma, serum or CSF sample, and maintaining the samples so as to minimize protein degradation and preferably coagulation in the case of plasma. Plasma samples are preferably maintained in the presence of protease inhibitors and/or coagulation-retarding compositions and at low temperatures (e.g., 4 C).

In a further aspect, the invention comprises recording information about a clinical or biological characteristic of a biological sample. Preferably, a biological sample is tested for the presence of aberrant levels of one or more abundant proteins, or tested to assess protein integrity or degradation. The method preferably involves obtaining clinical data such as a medical history, a family history, or biological and lab data test results. The information about a donor's medical history or information derived therefrom, or information about a clinical or biological

characteristic of a sample is preferably associated with a unique identifier for the donor or for a biological sample obtained from said donor. As will be appreciated, the information about a donor or biological sample is preferably stored in a database, preferably a computerized database.

It will be appreciated that the methods and systems of the present invention can be embodied in a variety of formats. In one exemplary method, populations including diseased and control groups are defined in order to minimize differences in plasma proteins unrelated to the presence or absence of the specific disease. Populations are defined by clinical and biological exclusion and inclusion criteria including: definition of the specific disease, age, concomitant diseases, and other factors as desired. In addition to statistical criteria, sample size of the populations may be defined according to, e.g., the volume and concentration of a pooled sample necessary to attain a preset level of sensitivity in the research method under consideration. Each donor enrolled will have standard laboratory measures (including clinical chemistry, complete blood count, coagulation tests, serology) assayed in a core laboratory in order to provide a distribution of standard clinical laboratory measures and risk factors among individual donors in the disease and control population. Some donors will be excluded from the pooling if the laboratory values are not in the normal standard ranges. Donors will be matched for age, ethnic group and other applicable criteria such as baseline biological parameters and medical evaluation. In preferred embodiments, assessment of subjects for biological parameters and medical evaluation involves assessing for a trait related to a disease to be studied in the research method.

BRIEF DESCRIPTION OF THE FIGURES

Figure 1 illustrates the workflow of the steps constituting the method of the invention. Boxes with dotted lines represent optional steps.

Figure 2 shows a system in which methods and systems consistent with the present invention may be implemented.

Figure 3 shows the components of a desktop or a server computer of the system of Figure 2.

Figure 4 illustrates the laboratory data values for the subjects evaluated for the research method described in Example 1.

DETAILED DESCRIPTION OF THE INVENTION

Definitions

As used herein, the terms "sample" and "biological sample" are used interchangeably, and include material derived from a mammalian subject, e.g., human. In preferred embodiments, a sample is a biofluid sample derived from a human or animal. Such biofluid samples include but are not limited to blood, plasma, serum, serum derivatives, bile, phlegm, saliva, sweat, urine, amniotic fluid, synovial fluid and cerebrospinal fluid (CSF), such as lumbar or intraventricular CSF.

The terms "detectable trait" and "trait" are used interchangeably herein and refer to any visible, detectable or otherwise measurable property of an organism including but not limited to: genetic makeup; symptoms of or susceptibility to a disease; clinical state; protein or enzyme levels; or use of a prescribed or non prescribed drug or other substance. The terms "detectable trait" and "trait" also include generally any "phenotype", including symptoms of or susceptibility to a disease; an individual's response to an agent, drug, or treatment acting on a disease; or symptoms of or susceptibility to side effects to an agent acting on a disease.

As used herein, an abundant protein constitutes greater than about 5% (w/w), more preferably greater than about 20% (w/w) of total protein in the sample.

The term "collection establishment" as used herein refers to any sample collecting organization. Collection establishments are typically regulated by the Food and Drug Administration or other regulatory agencies. A collection establishment can be either an independent entity or owned by a contractor. The term "contractor" as used herein refers to an entity that acts by contract as an intermediary between collection establishments and end-users. A contractor may be an end-user. The contractor queries collection establishments for individuals or samples that meet the criteria established by an end-user and arranges the supply of contact information or of those samples to the end-user. The contractor also provides end-users with access to databases according to the invention. The contractor may audit end-users to ensure the proper use of the information or samples by the end-user under the terms of the contract. The contractor's role as an intermediary does not preclude the contractor from undertaking additional functions of the invention including, but not limited to, sample preparation, storage, and shipping, sample quality analysis, etc.

The term "donor" as used herein means an individual who offers to donate or sell a biological sample. A donor typically provides a sample to a collection establishment. Donors fitting particular profiles also may be identified through partnerships with physicians, medical centers, and other health care providers. The sample provided by a donor may or may not be included in a particular study protocol or research method. Thus, the term donor encompasses potential donors, potential subjects, and subjects.

Some of the traits that are considered according to the methods of the invention are traits that affect or may affect the protein quality of a biological sample. "Affecting protein quality" of a sample refers to abnormal synthesis, processing, modification, elimination or degradation of one or more proteins, or abnormally high levels of one or more high-abundance proteins. Since it is often desirable to have biological samples that are matched and homogenous for factors other than those under study, it can be advantageous to eliminate samples that are suspected to have significantly different protein expression patterns where such expression is known to be unrelated to the condition or trait being studied. "Affecting protein quality" thus also includes factors such as differences in protein levels (resulting, for example, from differences in expression or stability), differences in protein modifications (e.g., phosphorylation, or other posttranslational modifications), and differences in protein form (e.g., splice variants, proteolytic targets, isoforms, etc.). These factors may be compared to other donors in the study protocol or under consideration for pooling, particularly where the factor(s) are related to a trait or condition other than the particular condition under study.

The term "end-user" as used herein means any entity that requests the unique identifiers of donors for assessment. End-users also include any entity that orders biological samples from a collection establishment for proteomics purposes and any entity that uses the database of the invention.

The term "longitudinal" as used herein means obtained over a period of time. When the term "longitudinal" is applied to an individual or group of individuals, the period of time, in general, extends from an individual's first to last sample donations. The last sample donation may occur, for example, when the individual develops a disease, when the individual begins treatment of a disease, when the individual stops treatment of a disease, or upon the death of the individual. When the term "longitudinal" is applied to a sample or to information, the period of time may extend beyond the death of the individual from whom the sample or information was gathered.

The present invention provides methods for obtaining biological samples in order to minimize difficulties in sample preparation/separation, and to provide samples having better and more reproducible quality.

Analysis of a proteome may allow the identification of proteins useful as protein therapeutics, as biological targets for intervention via an interacting molecule, or as biomarkers for the characterization of tissue and diagnosis of disease. Biochemical markers, for example, can be identified by analyzing tissue or body samples from a subject, preferably a mammal, with the disease of interest and then comparing the results of the analysis with those obtained from a subject without the disease. One successful approach using two-dimensional gel electrophoresis has led to the identification of a variety of marker proteins that are present at a significantly different concentrations in tissue or body fluid samples of a diseased mammal relative to a normal mammal. See, for example, Partin et al. (1993) Cancer Res. 53:744-746 which describes the identification of prostate cancer markers and Getzenberg et al. (1996) Cancer Res. 5 6:1690-1694, which describes the identification of bladder cancer markers.

Currently, a user may select among a number of fractionation strategies for fractionating biomolecules to be analyzed by proteomics. A widely used method is two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). The preparation and solubilization of any protein mixture for subsequent 2D-PAGE separation is of major importance because it affects the overall performance of the technique. Preparation for 2D-PAGE analysis includes (a) solubilizing as many proteins as possible; (b) disrupting protein aggregates; and (c) achieving reproducible sample preparation.

Other non-gel electrophoresis-based methods are also available. It has generally been thought that a substitute to 2D-PAGE should resolve proteins as well and also allow the rapid identification of resolved proteins. Attempts directed at alternative methods have included one and two-dimensional (1D and 2D) chromatography methods using high performance liquid chromatography (HPLC), capillary isoelectric focusing (CIEF), capillary electrophoresis (CE) or microcapillary chromatography. In the simplest liquid chromatography based methods, non-PAGE means have included 1D chromatography and mass spectrometry (MS). In one example, CIEF and electrospray ionization (ESI) MS have been investigated (Tang et al, (1997) Anal. Chem. 69: 3177-3182). Two-dimensional chromatographic methods linked to MS or MS-MS have been developed as well.

Several examples of systems for protein identification using peptide mapping have been described. In one example, Raida et al ((1999) J. Am. Soc. Mass Spectr. 10: 45-54), analyzing

peptides from human plasma filtrate by cation-exchange chromatography followed by reverse-phase HPLC, detected over 3000 distinct peptide masses but were able to determine the identity of only relatively few peptides. Wall et al. ((2000) Anal. Chem. 72: 1099-1111) examined a human erythroleukemia cell lysate using IEF apparatus and separating fractions obtained thereby using reverse phase HPLC column. Opiteck et al provide a system using a cation exchange column eluted in stepwise fashion onto a reverse phase column (Opiteck et al, (1997) Anal. Chem 69: 1518-1424). In another system, Opiteck et al coupled size exclusion and reverse phase HPLC (Opiteck et al, (1998) Anal. Biochem. 258: 349-361).

Other methods have focused on coupling chromatographic means to MS-MS. In one system, E. coli proteins were fractionated by anion exchange HPLC and portions were digested with trypsin and processed on a reverse phase microcolumn HPLC (Link et al, (1997) Int. J. Mass. Spectrom. Ion Proc. 160: 303-316). In another example, a mixture from Saccharomyces cerevisiae ribosomes was loaded onto reverse phase and eluted onto a cation exchange column from which peptides were separated and sprayed into an ESI tandem mass spectrometer (Tong et al, (1999) Anal. Chem. 71: 2270-2278). In a further system, a peptide mixture from Saccharomyces cerevisiae was loaded onto a biphasic 2D column packed with cation exchange and reverse phase materials and eluted onto an ESI MS-MS (Link et al. (1999) Nat. Biotechnol. 17: 676-682). Other methods and further details are provided in "Protein liquid chromatography", M. Kastner Ed., 2000, Elsevier, disclosures of which are incorporated herein by reference.

Various improvements to known fractionation systems have also been made. For example, improvements have focused on: i) enrichment of low abundance proteins (subcellular and protein pre-fractionation) prior to their solubilization, ii) depletion of high abundance proteins (e.g. albumin or IgG when a plasma sample is used), iii) removal of interfering compounds (salt, DNA, lipids and proteases), iv) use of different solubilizing agents and finally v) uses of a number of preparation and solubilization methods for samples such as body fluids, cell and tissue samples. A preferred fractionation system of the invention attempts to improve on past results by combining many of these strategies. Advantageous methods for a fractionation protocol, such as depletion of high-abundance proteins, protein concentration methods, size fractionation and other chromatographic methods, may be used to accomplish the above-listed improvements. For example, salts and contaminants are removed during concentration and chromatographic methods.

Depletion of high abundance proteins may be accomplished by methods common in the art, for example, using commercially available reagents. Protein A and Protein G are typically

used to bind to and remove immunoglobulin proteins from serum, lymph, and other biological fluids. Alternatively, antibody-based methods can be used to remove specific high abundance proteins from a sample.

Exemplary chromatography methods include ion-exchange column chromatography; chromatography using silica gel or a cation-exchange resin such as DEAE; gel filtration using, for example, Sephadex G-75; protein A Sepharose columns to remove high abundance Ig polypeptides; and specific antibody columns. Further removal of contaminants may be accomplished by ethanol precipitation; reverse phase HPLC; chromatofocusing; SDS-PAGE; ammonium sulfate precipitation; ultrafiltration and dialysis techniques (see, for example, Scopes, R., *PROTEIN PURIFICATION*, Springer-Verlag, New York, N.Y., 1982). These methods may be used in conjunction with a protein concentration step, using for example, a commercially available centrifugal protein filter (Millipore® Bedford, MA or Nalge Nunc Rochester, NY).

Determining which biological samples to use and/or the composition of samples required for a particular research method (e.g. proteomics study) can be carried out according to known methods. Preferably, the methods of the invention comprise selecting a statistical method to analyze the results of the research method, and deriving therefrom the number of biological samples needed. A user may have, for example, constraints related to the fractionation method or a detection (e.g. mass spectrometry) method. For example, a minimum amount of protein sample is required due to the sensitivity of the MS device used for detecting proteins present in low concentrations. Another example constraint may be that a minimum number of samples is required for a sample pool in order to reduce statistical error from multiple runs of a fractionation procedure. Typically, a user will also take into consideration the number of samples needed in view of a particular study design, such as the minimum number of subjects required in order to associate a protein with a trait of interest. The number of samples required will depend on a number of factors, including, for example, whether expected differences (e.g. protein concentration in a disease and control) are large or small, level of significance to be used, and natural variability in measurements or diagnosis of donors. Preferably, before calculating a sample size, a user will define one or more desired criteria of study (e.g. a clinically important diagnostic factor) for which he intends to compare samples from trait-displaying and control donors. Criteria may be standards used in the art or criteria identified by the user, for example, as recommended by an Institutional Review Board (IRB). The user may also define a level of statistical significance to be used, and the power to be used in the study.

The user can typically define sample size using standard sample size software, for example that provided with "Sample Size Tables for Clinical Studies (2^{nd} Edition, 1997) by D.

Machin, M. Campbell, P. Fayers, and A. Pinol, Blackwell Sciences Ltd. Other examples of software are StatsDirect (http://www.statsdirect.com/) or Epi Info (available from the CDC, Atlanta, http://www.cdc.gov/epiinfo). Further information is also available in "Statistical Issues in Drug Development", Ed. S. Senn, John Wiley & Sons (1997); Epidemiology: Study Design and Data Analysis, Ed. M. Woodward, Chapman & Hall/CRC (1999); "Biometry: the Principles and Practice of Statistics in Biological Research", Ed. R. Sokai et al., W.H. Freeman Co. (1995); and "Basic and Clinical Biostatistics", Ed. B. Dawson and R. Trapp, Appleton & Lange (2000).

Traits affecting sample quality

Traits that affect or may affect the protein quality of a biological sample can be of various types. For example, a trait may affect protein quality by resulting in different synthesis, elimination, or degradation rates of a particular protein, or differences in the level of high-abundance proteins. A trait may also result in differences in a particular posttranslational modification, proteolytic product, or splice variant of one or more proteins. Since it is often desirable to have biological samples that are matched and homogenous for factors other than those under study, it can be advantageous to eliminate samples that are suspected to have significantly different protein expression where such expression is thought to be unrelated to and/or not influenced by the condition or trait being studied. Affecting protein quality thus also includes aberrant levels of proteins in comparison to other donors from which samples under consideration or for pooling are derived, particularly where aberrant levels of proteins are related to a condition other than a particular condition under study.

Because abnormally high levels of highly abundant proteins can interfere with purification steps used in e.g. proteomics studies, it is desirable to select against such conditions when selecting biological samples. Preferably, samples from donors having a condition resulting in abnormal expression of immunoglobulins are identified and eliminated. One example of disorders involving aberrant levels of high-abundance proteins is multiple myeloma.

Traits related to liver impairment

Impaired liver function often results in aberrant elimination of proteins from circulation and thus diminishes the quality of biological samples.

Hepatitis is a chronic liver disease, generally meaning "inflammation of the liver", and is used to describe diseases resulting in hepatocellular damage. Hepatitis is commonly caused by infections, toxic agents or autoimmune diseases. Obstructive hepatitis, however, is caused by

obstruction of the biliary tract. Viral hepatitis is the most common cause of acute and chronic hepatocellular damage. Diagnosis of hepatitis is based on clinical features and measurement of liver enzymes. Serum levels of the serum transaminases ALAT and ASAT rise rapidly during the early courses of hepatitis because of parenchymal liver disorders. Serum alkaline phosphatase and gamma-glutamyltransferase are elevated during the early cholestatic portion of the disease and remain elevated until the disease has resolved. Diagnosis of the type of viral hepatitis is accomplished by detection of the specific hepatitis antigen in serum during the prodromal phase of the illness.

In one aspect, specific subjects suffering from liver impairment are identified. Chronic liver disease is associated with aberrant protein levels: albumin is decreased and this is generally accompanied by an increase in the beta and gamma globulins as a result of production of IgG and IgM in chronic active hepatitis. The alpha1 fraction of the serum protein globulin is decreased in chronic liver disease. A decrease in plasma fibrinogen can also be noted.

In evaluating liver impairment, liver enzymes can be assessed using standard laboratory methods known in the art. For example, normal liver function can be considered as follows:

- Alkaline Phosphatase (ALP): standard values are 30 to 125 IU/L.
- Gamma-glutamyltransferase (GGT): standard values are 9 to 40 IU/L in male and 9 to 35 IU/L in female.
- ASAT (Aspartate aminotransferase): normal ranges of 14 to 50 IU/L in male and 11 to 42 IU/L in female.
- ALAT (Alanine aminotransferase): normal ranges of 12 to 50 IU/L in male and 9 to 42 IU/L in female.
- Lactate dehydrogenase (LDH): standard values are 125 to 240 IU/L.

In addition to liver enzymes, other hepatic analyses can be considered as well. In one example, total, conjugated, and unconjugated bilirubin are measured. Total bilirubin levels in adults with normal liver function is generally about 6.8 to about 25 micromol/L, conjugated bilirubin levels about 1.7 up to about 8.6 micromol/L. In another example, cholesterol is measured. Total cholesterol normal ranges are 3.3 to 5.2 mmol/L and HDL cholesterol 0.75 to 1.85 mmol/L in male and 0.91 to 2.21 mmol/L in female. Triglycerides normal range is < 2.28 mmol/L.

Impaired liver function can also be caused by pharmacological agents. Drug-induced hepatic damage can occur, for example, from the use of barbiturates, amiodarone, tricyclic

antidepressants, antiepileptics, isoniazid and acetominophen. Thus, it is desirable to eliminate samples derived from subjects using or being treated with such drugs.

Other traits include excessive alcohol use, and disorders caused thereby. Alcoholic liver disease is due to chronic excessive ingestion of alcohol and results in an increase in globulin levels with a decrease in the albumin fractions of the serum proteins. In another example, cirrhosis is a liver disorder characterized by loss of normal microscopic architecture with fibrosis. Cirrhosis has a variety of causes but is most commonly secondary to chronic alcohol abuse. In cirrhosis, varying degrees of hyperglobulinemia may occur.

Traits related to renal impairment

Impaired renal function in a donor is another trait, which can reduce the protein quality of a biological sample from that donor. Potential impairment of the renal function can be assessed by testing of glomerular function, testing of tubular function (concentration and dilution studies), and urinalysis.

In one aspect, specific donors suffering from acute glomerulonephritis are identified. Acute glomerulonephritis is an acute inflammation of the glomeruli, resulting in oliguria, hematuria, increased Blood Urea Nitrogen (BUN) and serum creatinine levels, decreased Glomerular Filtration Rate, edema formation and hypertension.

In another aspect, specific subjects suffering from acute renal failure, renal insufficiency or chronic renal failures are identified. Acute renal failure is an abrupt deterioration in renal function. It is usually accompanied by oliguria or anuria. It is also associated with varying degrees of proteinuria, hematuria and the presence of red blood cell casts and other casts in the urine. Causes of acute renal failure are circulatory (hypovolemia, cardiac failure), renal (acute tubular necrosis, for example) or postrenal (obstruction of lower urinary tract).

Renal insufficiency can be defined as a creatine clearance < 40 ml/mn according to the Cockcroft formula (see for example Myara I, Lahiani F, Cosson C, Duboust A, Moatti N, Estimated creatinine clearance by the formula of Gault and Cockcroft in renal transplantation, *Nephron* 1989;51(3):426-7).

Chronic renal failure is a clinical syndrome resulting from the progressive loss of renal function. Causes of chronic renal failure are primary glomerular diseases, renal vascular diseases, and inflammatory diseases, for example.

Identifiable traits related to disorders causing impairment of renal function include nephrotic syndrome. Nephrotic syndrome has been classically defined as a clinical entity

characterized by massive proteinuria, edema, hypoalbuminemia, hyperlipidemia and lipiduria. This syndrome is characterized by increased glomerular membrane permeability that results in massive proteinuria and excretion of fat bodies. The causes are various including association with various forms of glomerulonephritis, generalized disease processes (amyloidosis, carcinoma, systemic lupus erythematosus) and mechanical or circulating disorders (renal vein thrombosis or constrictive pericarditis). Yet further examples of disorders causing impairment of renal function include vascular diseases, for example hypertension and arteriolar disease.

Impaired renal function can also be assessed by measuring serum creatinine levels and urea levels. For serum creatine assessment, normal score ranges in adult male are 62 to 106 micromol/L and 35 to 88 micromol/L in female. For urea (BUN) assessment, normal ranges are 2.8 to 7.1 mmol/L.

Impaired renal function can also be caused by pharmacological agents, such as the use aminoglycosides and analgesics, or be caused by chronic heavy metal poisoning. Thus, it is desirable to eliminate samples derived from subjects so treated or poisoned.

Traits related to other dysfunctions

In other aspects, the methods of the invention involve eliminating or selecting against any other trait related to a disease or condition expected to cause major alteration in protein composition. For example, for traits expected to cause alteration in a plasma sample, it is possible to identify subjects suffering from multiple myeloma, Waldenstrom's macroglobulinemia or proteinuria. Diagnosis can be carried out according to known methods.

Multiple myeloma is a malignant proliferation of plasma cells derived from a single clone. Bone pain is the most common symptom in myeloma affecting 70 percent of patients. The classic triad of myeloma is marrow plasmacytosis (>10 percent), lytic bone lesions and a serum and/or urine M component.

Waldenstrom's macroglobulinemia is a malignancy of lymphoplasmacytoid cells that secrete IgM.

Proteinuria is the presence of protein in urine and includes glomerulopathies such as diabetic nephropathy with microalbuminuria so named because of the abnormal albumin excretion of 30 to 300 mg per 24 hour, and tubular disorders.

In another example, subjects suffering from diabetes are identified and not included in the research method. In yet another example, subjects suffering from cancer are identified and not included in the research method.

Systems and methods

Systems and methods consistent with the present invention provides means that allow donors or samples for proteomics analysis to be selected based on factors which affect the quality of the sample for subsequent proteomics analysis. The methods enable access to a group of individuals and/or samples therefrom, whose medical data (for example, demographic characteristics, genetic markers, biochemical markers, family histories, and medical histories) make them attractive candidates for proteomics studies.

Such systems may be implemented in a simple selection scheme wherein previously collected samples are selected or eliminated from consideration based on factors that affect or potentially affect protein quality. The systems and methods can also be implemented in a network of non-profit and/or for-profit organizations and partners that have not traditionally been involved in this area of proteomics research. For example, a network of collection establishments refers donors into specific proteomic studies and collects blood samples and information from donors under Institutional Review Board (IRB) approved procedures and informed consents.

In one aspect, the problem of providing pooled protein samples of improved protein quality and representative of the average protein sample is addressed by selecting donor samples to be pooled from a large, diverse population of individuals with well-documented medical histories and detailed clinical and biological profiles. Donor subjects may be recruited from a variety of sources, including, but not limited to, individuals with specific diseases identified through partnerships with physicians and medical centers.

Another implementation consistent with the present invention provides proteomics researchers with access to a set of biological samples or a pooled sample, including, but not limited to, whole blood, plasma, and serum. Preferably, the biological sample donor(s) is (are) free of a trait affecting quality of protein in the biological sample.

Yet another implementation consistent with the present invention comprises a longitudinal database in which medical and demographic information for each donor, whether obtained through a collection establishment or through partnerships, is stored over time. The samples collected from an individual over time also are stored and may be retrieved by accessing a longitudinal database of samples. For example, a donor's first sample donation through either the development or amelioration of disease or death of that individual may be stored.

In other embodiments, as the database comprises vast amounts of data from large numbers of individuals, researchers are able to query the database so as to identify further factors or traits, which may affect quality of a protein sample. For instance, if certain protein quality characteristics of particular samples are determined experimentally, data from samples can be queried for unexpected correlations of certain protein quality characteristics with traits such as disease phenotypes or genetic background.

Overview of System Components and Operation

In its simplest embodiments, the invention can be carried out by manual implementation. However, preferably, the invention is carried out in a computer-assisted implementation wherein unique sample or donor identifiers and medical or other data associated with a donor and/or sample are stored in a database. The database can be queried using a computer implemented querying means.

In further embodiments, the implementation may comprise a network, preferably a computer network. For example, the implementation may comprise a contractor, a network of collection establishments and, optionally, partners, and end-users. Systems consistent with the present invention enable end-users, for example, researchers conducting a protein or proteomic analysis, to select samples or donors to be included in the study. Suitable subjects will vary from study to study and may be selected based on criteria such as age, sex, ethnicity, race or disease affection. The skilled artisan will recognize, of course, that many other selection criteria also may be appropriately applied depending on the particular requirements of the study.

Donor and sample Information

Multiple collection establishments are intake sites for prospective donors, optionally in collaboration with one or more partners. The collection establishments obtain informed consent from prospective donors in compliance with Institutional Review Board-approved procedures permitting, for example, the use of donated tissue samples in biomedical research and/or the release of donor contact information to end-users seeking additional information on the individual.

The collection establishments also collect donor information. Donor information, and more particularly "medical data" can include for example any trait displayed by the donor, demographic information, family histories, and medical histories, and, optionally, protein quality

analysis or clinical chemistry analyses on donor samples. Medical data may also generally include any information which may have an impact on or be indicative of a subject's health, susceptibility to disease, physical or mental condition, etc. In preferred aspects, information permitting the identification of conditions or traits which may affect protein quality is collected. Most preferably medical history is collected. Medical history comprises, for example, information relating to whether a patient is suffering or suspected of suffering from a medical disease or disorder and/or genotyping information, proteomics or any other suitable medical diagnostics information. If desired, genotyping and proteomics information (e.g. direct information on protein quality) can be can be carried out on the samples obtained from the donors. Preferably, a portion of a donor's samples is set aside for analysis. Donor DNA and RNA can be extracted from the sample using methods, either manual or automated, known to those skilled in the art.

Demographic information may include, for example, donor name, donor social security number, donor contact information, donor birth date, etc. Medical history information may include exposure to an infectious agent, such as exposure to hepatitis, HIV, or any disease or familial history of disease.

Additional information of use to the end-user may be collected, either prospectively or retrospectively. One skilled in the art will readily recognize that the nature of the donor information requested is dictated by the requirements of the study in which the donated sample is to be used.

The information collected is gathered by any available mechanism, including, but not limited to, confidential, health and habits questionnaire; self-executed forms; or by direct entry into a computerized database (e.g., via a personal computer terminal or a hand-held device). The information collected from prospective donors may be generally the same as is collected by collection establishments and is maintained in confidence.

Genomic and proteomic or protein quality information can be collected based on standard diagnostic testing. This usually involves taking a biological sample from a donor and performing a test. It will be appreciated that the testing can be carried out directly on the sample which is being considered for use in the proteomics project or for protein analysis in the present method. The proteomics and genomics information may indicate that a donor has or is at risk of developing a medical disorder. The testing may also serve to directly characterize protein quality of a sample, for example by detecting levels of selected abundant proteins, or the structural integrity of selected proteins. Thus, proteomics information may also be used to directly identify

samples with diminished quality of the proteins in the sample. Protein qualities may be assessed by methods common to the art, such as non-denaturing gel electrophoresis; SDS-PAGE (e.g., one-dimensional or two-dimensional SDS-PAGE); Western blotting (e.g., to check for the quantity and/or quality of a particular protein); chromatographic methods (including FPLC and HPLC); mass spectrometry; and protein chips (for example, according to US Patent 6225047, disclosure of which is incorporated by reference in its entirety). Useful protocols for such methods are disclosed in Sambrook, et al. Molecular Cloning. A Laboratory Manual. (1989) 2nd, ed., Cold Spring Harbor Laboratory, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.; "Ion Exchange Chromatography", H. Roos, in "Protein liquid chromatography", M. Kastner Ed., 2000, Elsevier; WO95/252819; U.S. Patents 5538897, 5869240, 5572259, and 5696376; and Yates, J. Mass Spec., 33:1 (1998); disclosures of which are incorporated by reference herein in their entireties.

Thus, optionally, information from other tests associated with each sample (e.g., nucleic acid based tests to detect Single Nucleotide Polymorphisms or to monitor changes in gene expression or proteomic tests to detect aberrant protein expression or changes in the posttranslational modifications) are performed on each sample. Such diagnostic tests may be carried out either at the time the sample is acquired or retrospectively, for example to search for changes in DNA sequence, RNA expression, or protein activity that are associated with a research method result, a later-arising disease, or changes in disease severity.

Any suitable unique identifier can be used to refer to an individual and/or a sample. To protect the identity of donors, the invention may use alphanumeric strings, rather than names, to identify each donor. Such strings may be assigned by the contractor, the collection establishment, or the end-user. The collection establishment may assign unique, confidential identification numbers to donors. The collection establishment may also assign a unique, confidential identification number to each sample collected from a donor. In implementation consistent with the invention, these numbers are used to identify sample and donor information.

A biological sample, preferably a blood sample, can be obtained from a donor at any time by any member of the network, provided that ethical principles are respected. The samples are collected and stored according to specially adapted protocols in order to minimize deterioration of the protein quality of the sample. Protocols for collecting, maintaining and storing samples which follow regulatory standards are well known in the art and further discussed herein.

In one aspect, the invention involves first screening a donor for selection, and depending on the outcome of the screening, proceeding to obtain a biological sample from this donor. Thus, the prospective donor is then selected or eliminated based on the screening results. "Screening", as used herein, refers to determining whether a sample or donor has at least one characteristic to be used as a criteria for inclusion in a study protocol or research method. If not selected for inclusion in the study, the individual can be classified as a deferred donor. As described, selecting or eliminating a donor can be done manually or using via a computer and/or database-assisted implementation. For example, information about a prospective donor is collected without necessarily collecting a sample, and a query can be run in order to identify individuals based on any screening criteria.

In another aspect, a biological sample is collected prior to the screening of the donor (or a sample there from) for selection. In this aspect, the method involves selecting or eliminating the sample for inclusion in a study protocol, research method, or from inclusion in a pooled sample. Thus, a sample can be selected or eliminated based on the screening results. It will be appreciated that screening results can be obtained from the biological sample itself. Thus a sample can be experimentally characterized with respect to genetic information, and more preferably protein quality. If not selected for inclusion in the study, the individual or sample can be classified as a deferred donor or sample. It will be appreciated that where a sample has been collected prior to the selection or elimination of a donor from consideration, the identifier for the donor may be provided in the form of an identifier for the biological sample obtained there from. Again, selecting or eliminating a donor or sample can be done manually or using a computer and/or database-assisted implementation, as further described herein.

Method for Identifying Samples or Individuals

Once donors have been identified, related information has been collected, and optionally, biological samples have been collected, the method of the present invention can be used to select the samples or donors for inclusion in a proteomics research method. The included samples will be analyzed for protein content or complement. Preferably, the sample will be included in or eliminated from a sample-pooling step, where samples obtained from a plurality of donors are combined as a single composition.

The selection process is typically implemented via query of a database containing the aforementioned information. Screening criteria used to query the database are chosen by the database user. The criteria include at least one characteristic associated with the quality of a

protein sample. Typically, a characteristic is a medical trait or condition known to be associated with diminished quality of a protein sample. Exemplary characteristics include aberrant levels of abundant plasma proteins such as immunoglobulins and impaired renal or liver function.

In other examples, criteria are experimentally determined characteristics of a biological sample collected from a donor. For example, a biological sample may be subjected to a testing step in order to directly identify at least one characteristic of the sample (e.g., levels of abundant plasma proteins). Preferred characteristics for experimental determination relate to the protein quality of a sample.

Obtaining and processing biological samples from donors

The sample type collected may be any suitable source of protein. Methods for preparation and analysis for a wide range of cells and tissues are known. Preferably, the sample is a body fluid or solubilized protein derived from a tissue. Preferred samples include blood, plasma, serum, sweat, tears, urine, peritoneal fluid, lymph, vaginal secretion, semen, spinal fluid, ascitic fluid, saliva, sputum, or breast milk.

When blood is collected, portions of each sample are stored as whole blood or as any fraction of whole blood (e.g., plasma, serum, lymphocytes, erythrocytes, etc.).

Donor samples are stored under standard conditions known in the art, preferably at a centralized depository maintained by the contractor, although storage at multiple sites, which may be maintained by third parties, is consistent with the invention. In one embodiment, stored samples are bar coded with unique identifiers to facilitate their identification and retrieval from storage. The facility for sample handling and storage may include a system for robotic handling and retrieval of individual samples.

Once a set or subset of biological samples have been selected based on the screening criteria (preferably trait(s) related to protein quality), the selected samples can be combined for a proteomic study. Combining multiple samples in a single container (e.g. as a composition comprising biological sample from a plurality of individuals) is referred to as the "pooling" of samples.

Maintaining quality of samples

In order to maintain the quality of the protein in the biological sample, the sample is obtained and stored using known methods for maintaining a sample. In the case of blood or

plasma, anticoagulant and antiprotease compositions are added. When the sample is blood, blood cell activation, aggregation and adhesion inhibitors are added. Commonly used anticoagulants fall into two general classes: the thrombin inhibitors and the calcium chelators. Of the thrombin inhibitors, heparin is the most commonly used. It is a known inhibitor of acid phosphatase, lactate dehydrogenase, beta-hydroxybutyrate dehydrogenase, glutamyl transferase, creatine kinase, and restriction endonucleases. Of the calcium chelators, ethylenediaminetetraacetic acid (EDTA), sodium citrate and oxalate salts are commonly used. Further anticoagulants are provided in WO 95/14788, the disclosure of which is incorporated herein. Commonly used and commercially available antiproteases include serine protease inhibitors (e.g., reversible and irreversible thrombin inactivators, Xa factors), and inhibitors of cysteine proteases, calpain proteases and metalloproteases. Blood cell activation, aggregation and adhesion inhibitors include for example platelet activation inhibitors, platelet-platelet interaction inhibitors, phospholipid-binding inhibitors, and fibrinogen inhibitors. In general, the anticoagulation and antiprotease mechanisms are selected so as not to interfere with methods for detecting proteins in the biological sample.

In general, a blood sample will be introduced into a means for receiving blood, such as a syringe, capsule, or blood bag, which has blood-contacting surfaces that are coated with anticoagulant compositions. Anticoagulants as well as antiprotease compounds can also be added to the blood or plasma sample itself. Preferably, a combination of anticoagulants and/or antiprotease compounds is used. Exemplary cocktails comprise at least two serine protease inhibitors, blood coagulation-retardant compounds, and blood cell activation, aggregation and adhesion inhibitors.

It will be appreciated that any known procedures can be used to process and maintain biological samples. Several examples of sample preparation methods for use in proteomic analyses are provided in Sanchez, J. C. Practical aspects of 2-DE: Sample preparation and solubilization ABRF '98, SanDiego (1998), the disclosure of which is incorporated herein by reference.

Plasma and serum samples

In one example, plasma samples can be prepared according to the methods of Anderson and Anderson (Proc. Natl. Acad. Sci. USA 1977, 74, 5421-5425) or of Golaz *et al.* (Electrophoresis 1993, 14, 1223-1231), the disclosure of which are incorporated herein by reference. For example, venipuncture blood is collected in a syringe or tube containing a suitable anticoagulant such as heparin or EDTA. After blood collection, the syringe or tube is

immediately placed in an ice bath and brought to the laboratory for analysis. Upon arrival in the laboratory the blood specimen is centrifuged immediately at 2000 g for 10 min at 5 C to avoid haemolysis, and decanted. Then, the sample can be either processed immediately or stored at -70 C until analysis.

In another example for obtaining serum samples, venipuncture blood is collected in a sterile tube. After blood collection, the tube is immediately brought to the laboratory for analysis. Upon arrival in the laboratory, the blood specimen is allowed to clot for 30 min at room temperature. It is then centrifuged at 2000 g for 10 min and decanted. The sample can then either be processed immediately or stored at -70 C until analysis.

Plasma and serum are preferably solubilized. For example, an aliquot of 6.3 µl of human plasma/serum is mixed with 10 µl of buffer A (10% (w/v) SDS, 2.3% (w/v) DTT). The sample is heated to 95 C for 5 min and then diluted to 500 µl with ISO buffer (9M urea, 4% CHAPS (w/v), 35mM Tris, 65mM DTT, and a trace of bromophenol blue), and mixed by vortexing. A suitable volume of plasma/serum can then be introduced to a separation device, such as a 2-D gel electrophoresis device. For example, 60 microl (45 microg) of the final diluted plasma/serum sample are loaded onto the first dimension of separation. The remaining sample can be stored for further analysis at -70 C.

CSF samples

In another example, cerebrospinal fluid (CSF) can be prepared for analysis. CSF samples have been studied by Switzer, R. C. et al, Anal. Biochem. 1979, 98, 231-237; and Goldman, D. et al, Clin. Chem. 1980, 26, 1317-1322, disclosures of which are incorporated herein by reference. CSF can be collected by lumbar puncture in a sterile container, and immediately placed in an ice bath and brought to the laboratory for analysis. The CSF specimen must be centrifuged to remove circulating cells at 2000 g for 10 min at 5 C. The sample can then either be processed immediately or stored at -70 C until analysis.

Proteins must be solubilized for loading onto a 2D gel electrophoresis separation device. For example, an aliquot of 300 µl of human CSF is mixed with 600 µl of ice-cold acetone and centrifuged at 10000 g at 4 C for 10 minutes. The supernatant is discarded, and the pellet is mixed with 10 µl of buffer A. The sample is heated to 95 C for 5 min, diluted to 60 µl with ISO buffer, and mixed by vortexing. The final diluted CSF sample can then be loaded (45 µg) onto the first dimension of separation. The remaining sample can be stored for further analysis at -70 C.

Urine samples

In a further example, urine sample is prepared. Urine sample preparation, solubilization, and 2-D PAGE maps have been described by Anderson *et al.* (Clin. Chem. 1979, 25, 1199-1210) and Edwards *et al.* (Clin. Chem. 1982, 28, 941-948), the disclosures of which are incorporated herein by reference. Urine contains proteins in trace amounts (100 mg/L) and thus needs to be concentrated with concomitant salt removal prior to 2-D PAGE analysis. Either a) Morning specimen, b) Random spot specimen, or c) 24 h specimen methods are used to collect urine samples. The specimens are collected in a sterile container and then centrifuged at 2000 g for 10 min to remove cells, casts and crystals. 0.1 mg of sodium azide is added per liter of urine in order to inhibit bacterial growth. The sample can then be either processed immediately or stored at -70 C. In order to prepare the sample for introduction to a separation device, 500 µl of urine is pipetted into the sample reservoir of a Microcon-10® microconcentrator. The sample is centrifuged at 14000 g for 30 min. 490 µl of MilliQ water is added to the retentate and centrifuged again at 14000 g for 30 min. The assembly is removed from centrifuge, and the sample reservoir placed upside down in a new tube and centrifuged at 1000 g for 1 min. Between 10 and 20 µl of concentrated and desalted urine will be collected. An aliquot of 10 µl of concentrated urine are then mixed with 10 µl of buffer A. The sample is heated to 95 C for 5 min and diluted to 60 µl with ISO buffer, and then loaded (45 µg) onto the first dimension of separation. The remaining sample can be stored for further analysis at -70 C.

Sample Pooling

Samples can be maintained and used in a study (e.g. the protein complement analyzed) individually or as pools of samples. Biological samples may be pooled at any desired moment. In one aspect, samples may be pooled, and stored as a pooled sample, or more preferably treated, frozen and stored upon collection and pooled shortly before protein analysis. As known in the art, analysis typically involves separation or fractionation of proteins in the biological sample, and subsequently the identification of proteins using for example mass spectrometry methods (MS). Separation and fractionation can be achieved using known liquid chromatography or 2D gel electrophoresis methods.

In a preferred embodiment, aliquots of the biological samples are retained un-pooled (e.g. as individual samples). For example, half of the sample volume or weight may be retained frozen un-pooled. In another non-limiting example, a tenth of the sample volume or weight may be retained frozen un-pooled. These retained samples allow the end-user conducting the proteomics

study to assay each of the individual samples for a given protein observed in the pooled sample. Such an assay may be an immunoassay (e.g. an Enzyme Linked Immuno Sorbent Assay (ELISA) as known in the art). Such a procedure allows the end-user conducting the proteomics study to test whether a given protein is equally distributed among samples from one pooled sample or if its distribution is limited to a small proportion of the samples, in which case its usefulness may be diminished.

Databases

In a preferred example, a database containing information about a donor or biological sample is provided. A request to identify biological samples originates with an end-user. The end-user provides desired characteristics or selection criteria. The end-user may wish to identify donors with specific characteristics affecting protein quality of a sample obtained therefrom. Based on those characteristics, the contractor formulates a query, which is designed to interrogate the database for samples obtained from donors satisfying the criteria. Alternatively, the query can be designed to select the inverse, that is, to interrogate the database for samples obtained from donors not satisfying the criteria. The query is used to interrogate a database, and records in that database that satisfy the query are identified as unique donor identifiers by the computer.

Donor information and data associated with samples (e.g., medical history, sample storage location, protein quality information, SNP or protein expression profile, etc.), collectively "information," may be stored using any method that permits high productivity, scalability, flexibility, accessibility, security, correctness and consistency of housed data, data granularity, and presentation. The storage system may be a computerized database. In one implementation consistent with the invention, the information is stored in a secure, computerized data warehouse system, accessible only by controlled passwords assigned to trained users. In general, collection establishments currently use this type of system for data storage. The data warehouse is designed using dimensional modeling, a logical design technique that seeks to present the data in a standard framework that is intuitive and allows for high-performance access. This type of modeling provides the optimal balance among critical factors such as productivity scalability, flexibility, accessibility, security, correctness and consistency of housed data, data granularity, and presentation.

In one implementation consistent with the invention, end-users provide the contractor with criteria through which the desired donors and samples may be identified. The contractor causes the donor data and sample information database or databases to be searched using queries

developed using the end-user supplied criteria. Standard query protocols are used, resulting in the data required for the end-user to be output. In general, a query tool set is selected that allows for services such as warehouse browsing, query management, standard reporting, access and security.

Database queries are performed by trained employees either of the contractor or of the collection establishments. Database queries may be performed by the contractor, by employees of the collection establishments, or by end-users, following protocols establishing confidentiality and proper security. The result of a query is the approach to an individual donor to participate in an end-user's research, the shipment of sample to the end-user, or the identification of desired proteomic/genomic information.

Network implementation

It will be appreciated that the present invention may be implemented in a software system, which is stored as executable instructions on a computer readable medium accessible either directly or through a network.

Figure 2 illustrates a conceptual diagram of a computer network (400) in which methods and systems consistent with the present invention may be implemented to permit users to query a database of donor and sample information. Computer network (400) comprises one or more small computers (such as desktop computers, 410, 420, and 425) and one or more large computers (such as Server A (412) and server B (422)). In general, small computers are "personal computers" or workstations and are the sites at which a human user operates the computer to make requests for data from other computers or servers on the network. Usually, the requested data resides in the large computers, but the size of a computer or the resources associated with it do not preclude the computer's acting as the home of a database. In one implementation consistent with the invention, Servers A and B are connected through a firewall (435), which permits secure access to information that identifies donors to authorized users. In another implementation consistent with the invention, Servers A and B are not connected by a network and patient information must be accessed directly from server B.

Desktop computer systems and server systems compatible with the invention includes conventional components, as shown in Figure 3, such as a processor (524), memory (525, e.g., RAM), a bus (526) which couples processor (524) and memory (525), a mass storage device (527, e.g., a magnetic hard disk or an optical storage disk) coupled to processor (524) and

memory (525) through an I/0 controller (528) and a network interface (529), such as a conventional modem or Ethernet card.

The distance between a server (412) and a desktop computer (410) may be very long, e.g., across continents, or very short, e.g., within the same building. When the distance is short, the network (400) is preferably a local area network (LAN). When the distance between server (412) and desktop computer (425) is long, the network (400) may, in fact, be a network of networks, such as the Internet. In traversing the network, the data may be transferred through several intermediate servers and many routing devices, such as bridges and routers. Proper security and flexibility of access will be employed to provide authorized access through commonly used interface technologies.

The software system of the present invention is, for example, stored as executable instructions on a computer readable medium on the desktop and server systems, such as mass storage device (527), or in memory (525). Access to the system described above is available on a single-use or on a multiple-use basis. Preferably, end-users contract with the contractor for continuing access to the system.

It will be appreciated that any suitable network implementation can be used as well. For example, as criteria for selecting or eliminating a donor or sample are selected, the selection process is typically implemented via query of a database containing the aforementioned information. For example, a query is sent to Server A, which comprises the database, over a communications network. Records in that database that satisfy the query are identified and output as unique donor identifiers by Server A.

In one implementation consistent with the invention, the name and contact information associated with each identifier also are stored in the database. In another implementation consistent with the invention, the name and contact information associated with each identifier are stored in a second database, which cross references the unique patient identifiers with the names and contact information of the corresponding individuals. The latter allows for blinding of contact information and possibly more information, such that the final entity conducting the proteomics experiment has only access to unique patient identifiers for reference, and does not have access to those patient details which are not needed for conducting the proteomics study according to the method of the invention. If needed, the final entity may request from the collection establishment additional information, e.g. on a given sample or set of samples which,

from a proteomics standpoint, deviates from the study group to which they belong, even though the information available to the final entity does not allow to predict such a deviation.

In one implementation consistent with the invention, the proteomics sample database and the second database are stored on Server A. In another implementation, the second database is stored on a separate Server B.

In implementations of the invention utilizing Server B, Server A may be either directly linked to Server B through a firewall or, alternatively, freestanding and without links to other components of the communications network.

Information is retrieved from Server B either through the communications network if a link is present in the system or manually if Server B is freestanding.

In general, the contractor or the collection establishment contacts individual identified and seeks permission to pass patient contact information on to the end-user. Alternatively, the patient information may be sent directly to the end-user, who then contacts the individuals identified or, alternately, further refines the query for resubmission to the contractor.

The foregoing description of implementations of the invention has been presented for purposes of illustration and description. It is not exhaustive and does not limit the invention to the precise form disclosed.

EXAMPLE: selection of patients and controls for a Coronary Artery Disease study:

In this example, the selection process for a matched set of male patient and control plasma samples for the proteomic study of coronary artery disease is described.

Patients screened at the Duke Cardiac Catheterization Laboratory and control subjects identified from the Duke Databank for Cardiovascular Diseases were selected with the following criteria:

Coronary artery disease patient population:

- inclusion criteria: age 35 to 65, male, and coronary artery stenosis of >50% in at least one major coronary artery.
- exclusion criteria:

- Acute myocardial infarction within one month
- Significant valvular heart disease
- NYHA Class III or IV heart failure
- Cigarette smoking > 2 packs per day
- Most recent known total cholesterol > 300 mg/dL or triglyceride > 400 mg/dL
- Significant anemia (hemoglobin less than 11.0 g per dl for female donors, and less than 12.0 g per dl for male donors)
- Hypotension (<90 mm Hg systolic or <50 mm Hg diastolic)
- *Diabetes mellitus, treated with oral hypoglycemic therapy or insulin and/or with end-organ damage (retinopathy, nephropathy, neuropathy)*
- *Any other disease/condition expected to cause major alteration in plasma protein composition*
- *Uncontrolled hypertension (>180 mm Hg systolic or >100 mm Hg diastolic), and/or with end-organ damage*
- *Renal insufficiency (creatinine clearance < 40 ml/min (Cockcroft formula))*
- *Active malignancy*

It can be seen that the last 5 exclusion criteria (in italics) are in particular used to exclude patients with conditions likely to yield abnormal plasma protein concentration and/or quality, abnormality which would not be related to the disease under study.

<u>Control population:</u>

- inclusion criteria: age 35 to 65, male, no coronary artery stenosis of ≥ 25% on cardiac catheterization within 2 years and normal left ventricular ejection fraction and normal regional wall motion.
- exclusion criteria:
  - Typical symptoms of angina, or any evidence of myocardial ischemia on stress testing, myocardial infarction or unstable angina
  - Any history of peripheral arterial or cerebrovascular disease, including significant claudication, stroke, transient ischemic attack, or significant vascular stenosis on noninvasive imaging or angiography
  - Any symptomatic heart failure
  - Significant valvular heart disease
  - NYHA Class III or IV heart failure
  - Cigarette smoking > 2 packs per day

- Most recent known total cholesterol > 300 mg/dL or triglyceride > 400 mg/dL
- Significant anemia (hemoglobin less than 11.0 g per dl for female donors, and less than 12.0 g per dl for male donors)
- Hypotension (<90 mm Hg systolic or <50 mm Hg diastolic)
- *Diabetes mellitus, treated with oral hypoglycemic therapy or insulin and/or with end-organ damage (retinopathy, nephropathy, neuropathy)*
- *Uncontrolled hypertension (>180 mm Hg systolic or >100 mm Hg diastolic), and/or with end-organ damage*
- *Renal insufficiency (creatinine clearance < 40 ml/min (Cockcroft formula))*
- *Active malignancy*
- *Any other disease/condition expected to cause major alteration in plasma protein composition*

Here again, the last 5 exclusion criteria (in italics) are in important for ensuring optimal plasma protein concentration and/or quality.

A total of 97 CAD male patients and 91 male control subjects selected using the above-defined criteria were further selected based on laboratory values obtained for the corresponding samples, as illustrated in Figure 4, to eliminate subjects with abnormal laboratory parameters. As can be seen in Figure 4, the parameters have been grouped into categories, related to impairment of the liver, impairment of the kidney, parameters related to the disease under study and other parameters.

General guidelines and information regarding analytes measurements as conducted herein are available, for example from the International Federation of Clinical Chemistry and Laboratory Medicine (http://www.ifcc.org).

Importantly, total protein concentration is taken into account as a measure of quality of the sample (Figure 4a). Additionally, serum protein electrophoresis (Agarose system Paragon / Sebia) was used to detect disorders such as gammapathies: abnormal synthesis of immunoglobulins would indeed disturb significantly the downstream proteomic study, and these can be detected to some extent by a qualitative study of electrophoresis results by one skilled in the art.

A final matching for age and ethnicity of the patient and control groups yielded a set of 53 CAD male patients and 53 control male subjects to be used in the proteomic study.

Modifications and variations are possible in light of the above teachings or may be acquired from practicing of the invention. For example, the described implementation includes software but the present invention may be implemented as a combination of hardware and software or in hardware alone. The invention may be implemented with both object-oriented and non object-oriented programming systems. The scope of the invention is defined by the claims and their equivalents.

CLAIMS

1. A method for selecting biological sample donors for inclusion in a research method, the method comprising the steps of:
   (a) selecting one or more trait(s) for the assessment of donors, wherein at least one of these traits is affected by the quality of the proteins in the donor biological sample and wherein said trait(s) is used as a criteria for inclusion of the donor biological sample in the research method;
   (b) providing a plurality of donors;
   (c) for each donor, assessing whether said donor displays the trait(s) selected in step (a); and
   (d) selecting donors for inclusion in the research method, wherein said donors are selected according to display of the trait(s) assessed in step (c).

2. A method for selecting biological sample donors for inclusion in a research method, the method comprising the steps of:
   (a) providing a plurality of donors;
   (b) obtaining a biological sample and medical data from each donor;
   (c) recording medical data from each donor into a database and correlating said data with said biological sample from each donor;;
   (d) selecting one or more trait(s) for the assessment of each donor, wherein at least one trait is related to the quality of the proteins in the donor biological sample and wherein said trait(s) is used as a criteria for inclusion of a donor biological sample in the research method;;
   (e) for each donor, assessing whether said donor displays the trait(s) selected in step (d);
   (f) selecting a donor for inclusion in the research method, wherein said donor is selected according to display of the trait(s) assessed in step (e).

3. The method of claims 1 or 2, further comprising selecting one or more biological sample type to be used in a research method.

4. The method of claims 1 to 3, further comprising selecting or providing one or more fractionation strategy(ies) for a selected biological sample type(s).

5. The method of claims 1 to 4, further comprising selecting a statistical method for the analysis of the results of the research method and deriving there from numbers of biological samples needed.

6. The method of claim 5, further comprising assessing whether the one or a plurality of donor(s) or prospective donor(s) selected for inclusion in the study protocol fulfills the number of biological samples needed, and optionally repeating any or more of the steps of the methods of Claims 1 or 5.

7. The method of claim 6, comprising repeating the methods of Claims 1 or 5, thereby selecting further biological samples for inclusion in a research method.

8. The method of any one of claims 1 to 7, comprising providing one or more biological sample(s) from at least a portion of said plurality of donors.

9. The method of any one of claims 1 to 7, comprising obtaining one or more biological sample(s) from at least a portion of said plurality of donors.

10. The method of any one of claims 1 to 9, further comprising performing at least one assay for a clinical or biological parameter to determine whether a subject displays at least one trait which may affect the quality of the protein in a donor's biological sample.

11. The method of claim 10, wherein said assay comprises detecting the presence of a protein in a donor's biological sample.

12. The method of claim 10, wherein the assay is a genotyping assay.

13. The method of any one of claims 1 to 12, further comprising combining a plurality of biological samples from the selected biological samples in a pooled biological sample.

14. The method according to any one of the above claims, wherein the selected trait is a trait relating to aberrant levels of a known protein.

15. The method according to any one of the above claims, wherein the selected trait is a trait relating to elevated levels of an abundant protein.

16. The method according to any one of the above claims, wherein a selected trait is a trait relating to impairment of elimination of proteins from circulation.

17. The method according to any one of the above claims, wherein a selected trait is a trait relating to impairment of liver function.

18. The method according to any one of the above claims, wherein a selected trait is a trait relating to impairment of renal function.

19. The method according to any one of the above claims, wherein a selected trait is a trait involving or potentially involving a substantial alteration in plasma protein composition.

20. The method according to any one of the above claims, wherein a selected trait is a trait involving elevated levels of an abundant protein.

21. The method according to any one of the above claims, wherein a selected trait is a trait involving elevated levels of immunoglobulins.

22. The method according to any one of the above claims, wherein a selected trait is a trait involving aberrant protein degradation.

23. The method of claim 10, wherein said assay is selected from the group consisting of: a liver impairment assay, an alkaline phosphatase (ALP) assay, a gamma-glutamyltransferase (GGT) assay, an AST (aspartate aminotransferase) assay, an ALT (alanine aminotransferase) assay, and a lactate dehydrogenase (LD) assay, an assay for bilirubin levels, an assay for cholesterol levels, an assay for triglyceride levels, an assay for serum creatine levels, and an assay for urea levels.

24. The method according to any one of the above claims, wherein the biological sample is a blood sample.

25. The method according to any one of the above claims, wherein the biological sample is a plasma sample.

26. The method of claims 21 or 22, wherein the sample is maintained so as to minimize protein degradation and coagulation.

27. The method of claims 21 to 23, wherein the sample is maintained in the presence of protease inhibitors and/or coagulation retarding compositions.

28. The method according to any one of the above claims, wherein a biological sample is obtained from at least 10 subjects.

29. The method according to any one of the above claims, wherein a biological sample is obtained from at least 50 subjects.

30. The method according to any one of the above claims, wherein a biological sample is obtained from at least 100 subjects.

31. The method according to claims 10 to 12 or 20, comprising recording information about a clinical or biological characteristic of a biological sample in a database.

32. The method according to claims 2 to 28, wherein the medical data comprises a subject's medical history.

33. The method of claims 2 to 29, wherein the medical data comprises a subject's family history.

34. The method of claims 2 to 30, wherein the medical data comprise clinical chemistry test results.

35. The method of claims 2 to 31, wherein the database is a computerized database.

36. The method of claim claims 2 to 32, further comprising querying the database for donors, prospective donors, or samples obtained therefrom displaying a selected trait.

37. The method of claim 1 or 2, further comprising the step of pooling the biological sample obtained from a selected donor with the biological samples obtained from other selected donors.

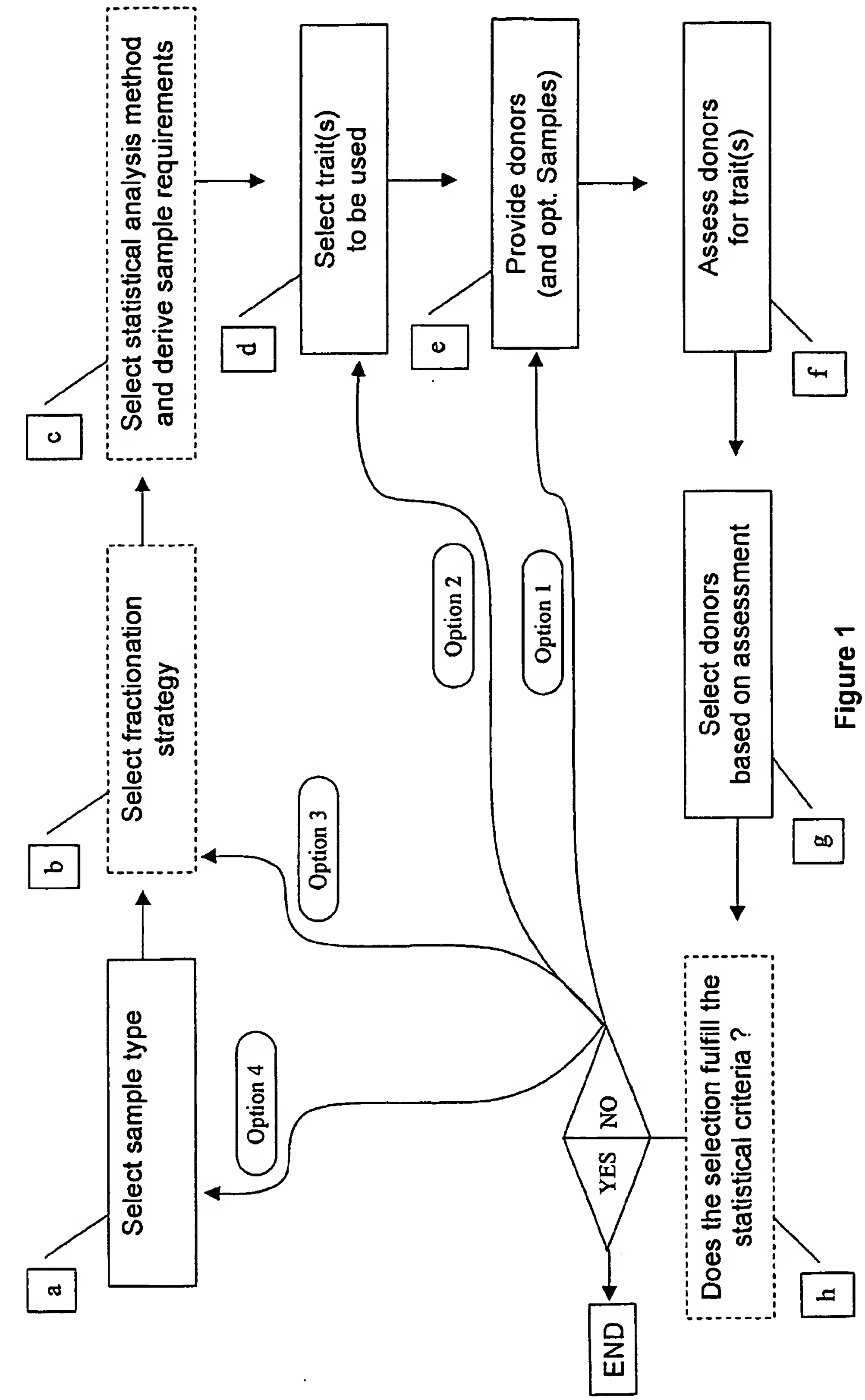
a
Select sample type
b
Select fractionation strategy
c
Select statistical analysis method and derive sample requirements
d
Select trait(s) to be used
e
Provide donors (and opt. Samples)
f
Assess donors for trait(s)
g
Select donors based on assessment
h
Does the selection fulfill the statistical criteria ?
YES
NO
END
Option 1
Option 2
Option 3
Option 4

Figure 1

Figure 2

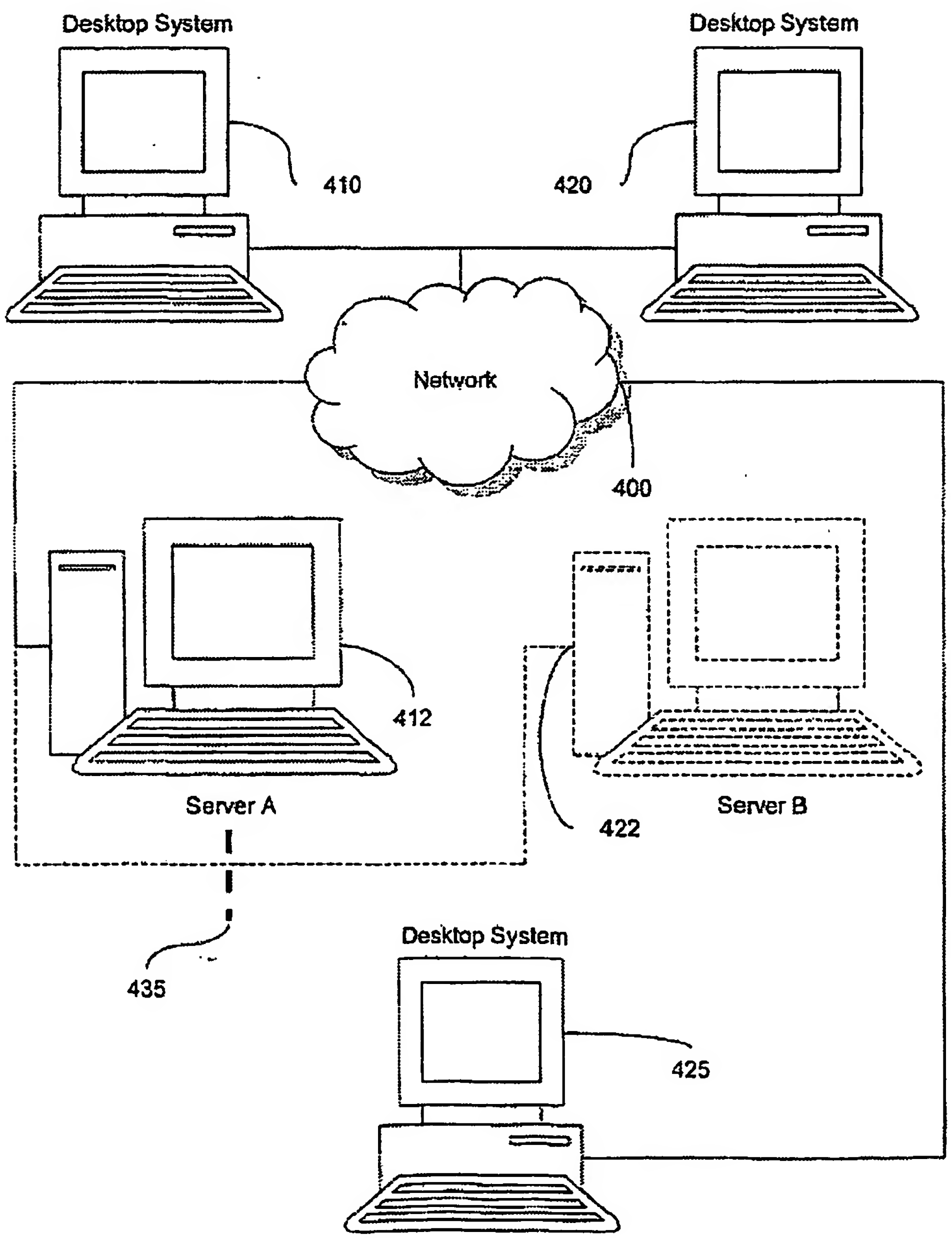
Desktop System
Desktop System
410
420
Network
400
412
Server A
422
Server B
435
Desktop System
425

Figure 3

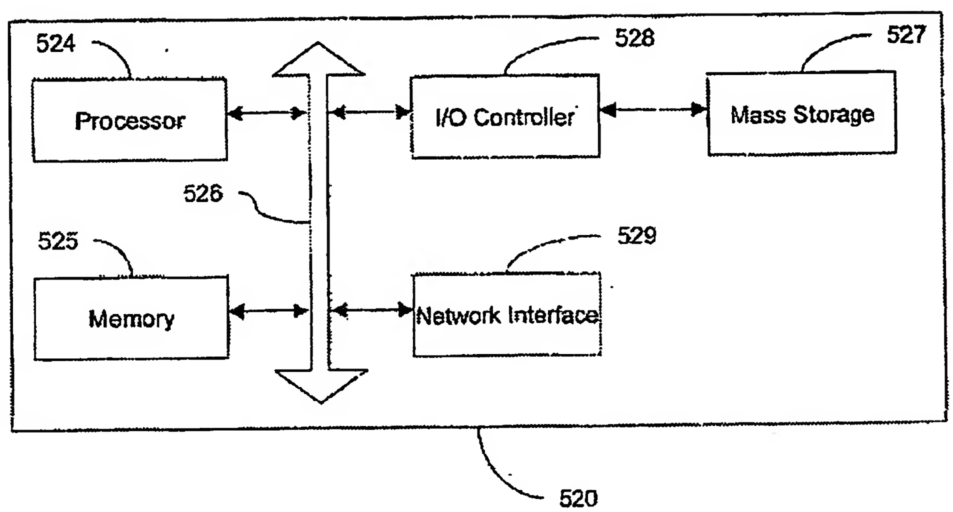
524
Processor
526
525
Memory
528
I/O Controller
527
Mass Storage
529
Network Interface
520

Figure 4a

Trait considered	Body Fluid	Units	low cut off*	high cut off*	Method	Remark
Liver impairment-related						
ALAT	plasma	U/l	12	50	IFCC (37 C)	
ASAT	plasma	U/l	14	50	IFCC (37 C)	
Albumin	plasma	g/l	35	48	Bromcresol purple	
Proteins	plasma	g/l	61	79	Biuret	
Alkaline phosphatase	plasma	U/l	30	125	IFCC (37 C)	
GGT	plasma	U/l	9	40	IFCC (37 C)	extended to 4 times the high cut off.
LDH	plasma	U/l	125	240	IFCC (37 C)	
Total Bilirubin	plasma	micromol /l	6.8	25	Diazo	
Direct Bilirubin	plasma	micromol /l	1.7	8.6	Diazo	
HDL cholesterol	plasma	mmol/l	0.75	1.85	Colorimetric direct method	
Total cholesterol	plasma	mmol/l	3.3	5.2	Colorimetric	extended to 2 times the high cut off.
Triglycerides	plasma	mmol/l	-	2.28	Colorimetric	extended to 2 times the high cut off.
Renal impairment-related						
Creatinine	plasma	micromol /l	62	106	Picric acid	
Urea	plasma	mmol/l	2.8	7.1	Conductimetry	
Renal impairment-related: urinary electrolytes						
Total calcium	plasma	mmol/l	2.20	2.52	Ion Selective Electrode	
Chorides	plasma	mmol/l	96	107	Ion Selective Electrode	
Phosphates	plasma	mmol/l	0.8	1.5	Phosphomolybdate	
Potassium	plasma	mmol/l	3.6	4.6	Ion Selective Electrode	
Sodium	plasma	mmol/l	136	144	Ion Selective Electrode	
Disease-related						
Total creatine kinase	plasma	U/l	47	222	IFCC (37 C)	extended to 2 times the high cut off.
Creatine kinase-MB	plasma	IU/l	2	14	Immuno-inhibition-enz.	extended to 2 times the high cut off if total CK within normal range
Troponin I	plasma	microg/l	-	0.09	Immuno-enzymo fluorometric	

* male subject only

Figure 4b

Trait considered	Body Fluid	Units	low cut off*	high cut off*	Method	Remark
			Others			
CRP	plasma	mg/l	0	10	Immunoturbidimetry	extended to 2 times the high cut off.
Fasting glucose	plasma	mmol/l	4.1	6.0	Oxygen electrode	extended to 2 times the high cut off if HgbA1 normal.
Glycated hemoglobin	plasma	%	3	6.	Homogeneous antibody complex	
Homocystein	plasma	micromol/l	5	15	Enzyme Immuno Assay / Imx Abbott	
Magnesium	plasma	mmol/l	0.65	1.05	Colorimetric	
			Thyroid			
Thyroid Stimulating Hormone	serum	mUI/l	0.4	4	Immunochemiluminometric assay / Immulite	

* male subject only